DOCUMENT RESUME

ED 326 312 PS 019 196

AUTHOR Meisels, Samuel J.; And Others
TITLE Testing, Tracking, and Retaining Young Children: An Analysis of Research and Social Policy.
SPONS AGENCY National Center for Education Statistics (ED), Washington, DC.
PUB DATE Dec 89
NOTE 46p.; A commissioned paper.
PUB TYPE Information Analyses (070) -- Guides - Non-Classroom Use (055)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Criteria; Criterion Referenced Tests; Early Childhood Education; Educational Practices; *Educational Testing; Grade Repetition; Guidelines; Kindergarten; Minority Group Children; Norm Referenced Tests; Racial Bias; Research Needs; *School Readiness Tests; *State Programs; *Test Selection; *Test Use; Track System (Education); *Young Children
IDENTIFIERS *Georgia

ABSTRACT

Many professionals are convinced that more testing, tracking, and retention are taking place in the early school years than ever before. They also believe that developmentally inappropriate modifications to curricula are being implemented. As a result of inappropriate use of standardized tests, disproportionate numbers of poor and minority children have been retained or placed in extra-year programs. This paper explores these issues and makes recommendations concerning uses of assessment data and alternatives to conventional testing practices. The report also discusses the large number of unready, at-risk children entering kindergarten. Sections of the text focus on: (1) issues and background on testing, tracking, and retention; (2) high stakes testing, i.e., the use of tests to make important decisions that immediately and directly affect those tested; (3) a rational perspective on tests and testing; (4) ways in which schools, teachers, and tests are failing minority children; (5) a rationale for testing young children, guidelines for deciding to use particular kinds of tests, characteristics of norm-referenced and criterion-referenced instruments, and criteria for selecting developmental screening instruments and school readiness tests; and (6) needed research about alternatives to standardized testing. Ninety-seven references are included. (RH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
X This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

ED326312

Testing, Tracking, and Retaining Young Children: An Analysis of Research and Social Policy

Samuel J. Meisels

Dorothy Steele

Kathleen Quinn

The University of Michigan

PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Samuel J. Meisels
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC).

PS 018396

A Commissioned Paper Prepared for the National Center for Education Statistics

December, 1989

Testing, Tracking, and Retaining Young Children: An Analysis of Research and Social Policy[1]

A. Introduction

Testing of school children has expanded dramatically during the past six years. The number of students taking the SAT has increased by 15 percent and nearly every state has translated its concerns for student achievement into formalized competency testing since the 1983 publication of "Children at Risk" by the U.S. Department of Education.

The drive for accountability has had a major impact on young children as well. In many school districts five and six year olds are expected to pass entry examinations before beginning kindergarten or first grade. Children who are judged "not ready" are retained in grade or are placed in extra-year programs before kindergarten or first grade.

The rise in early childhood testing has been accompanied by changes in the curriculum, these changes controlled by teachers' and administrators' perceptions of what children must learn in order to do well on the standardized tests they will encounter later in elementary school. As is the case with teachers of older students, kindergarten or first grade teachers experience pressure for their students to be successful on these tests and many of them alter their curricula to reflect the content of the tests. This situation results in a host of problems, including downward extension of academic curricula, rigidified content, homogeneous approaches to teaching, early tracking, and a reification of the concept of "readiness."

Many professionals are convinced that more testing, tracking, and retentions are taking place in the early years of school than ever before, and that developmentally inappropriate modifications to curricula are being implemented. Furthermore, the inappropriate use of standardized tests has resulted in disproportionate numbers of poor and minority children being retained or placed in extra-year programs.

---

1. This paper is based, in part, on material published in Meisels 1986, 1987a, 1987b, 1989a, and 1989b

The suggestion that large numbers of children entering public kindergarten may fail, or may not be prepared to benefit from the regular classroom program, is a cause for great concern. This paper will explore this issue and others mentioned above. Recommendations will be made concerning future uses of assessment data and alternatives to conventional testing practices.

B. Testing, Tracking, and Retention: Issues and Background

Most educators are extremely ambivalent about standardized tests. They love them, and they hate them; they adopt them, and they reject them; they need them, and they do not understand them. Moreover, whenever a new trend in education emerges, a national commission reports its recommendations, or a novel idea is introduced, standardized tests are usually mentioned as the preferred means of measurement, implementation, or evaluation.

Today we are witnessing an increasing commitment on the part of public schools to serve 4- and 5-year old children. Recent reports estimate that more than half of the states have enacted legislation providing for some form of early childhood education, and about one-quarter of the country's 15,000 school districts are offering formal instruction of some kind to four-year olds. Accompanying this rapid growth in early childhood programs is an inevitable controversy about testing. This paper will discuss several aspects of the current controversy.

It is becoming apparent that early childhood curricula are undergoing change and that standardized tests are either contributing to this change or are at least reinforcing it. Early childhood programs are focusing increasingly on narrowly construed academic objectives, behavioral compliance, abstract thinking, and one-dimensional teaching methods. Never before have we witnessed such a "downward extension" of traditional early elementary curriculum goals and methods into programs for four- and five-year olds (see Shepard & Smith, 1988). While early childhood programs have become more rigid, predictable, subject-matter oriented, and linear, they have also become more amenable to standardized testing. Indeed, academically-oriented early childhood curricula and group-administered standardized tests are a marriage made in heaven. Previously, when teachers and professionals were seeking to test and evaluate children in child-centered programs based on individuality goals, discovery learning, and

extensive opportunities for children's initiation and activity in the classroom, standardized tests were seen as a poor fit and were criticized as irrelevant and unhelpful (see Bryk, Meisels, & Markowitz, 1979; Carini, 1975; Hein, 1979). But today testing has become much more prevalent in public schools generally and in kindergarten programs in particular. For example, a recent survey of more than 300 school districts in Michigan reported that 83 percent of the districts annually test all children who are eligible for kindergarten (Riley et al., 1988). Reports indicate that more than half of the states require pre-kindergarten screening in compliance with Public Law 94-142 (Meisels, Harbin, Modigliani, & Olson, 1988), while testing of other kinds occurs in three-fourths of the states before children enter first grade (E. Fiske, 1988; Gnezda & Bolig, 1988).

The trend toward increased use of standardized testing with kindergarten and elementary school children has been well-documented in the press. During one four-month period in 1988 the Boston Globe published an article entitled, "Fears for a son going into a test-crazy world" (Yagelski, 1988), Time ran a story called, "Can Kids Flunk Kindergarten?" (Bowen, 1988), and the New York Times devoted five pages in their Spring Education supplement to "America's Test Mania" (E. Fiske, 1988). All of this attention is not simply a recent phenomenon. Madaus (1988) reports that, as measured by column inches in Education Index (a widely-used index of publications relating to educational issues), "attention devoted to testing has increased ten-fold in the last fifty years, rising from only 10 to 30 column inches in the 1930s and 1940s to well over 300 inches in the 1980s" (p. 84). Shepard (1989) has also commented on the rise in testing, saying that it is "running amok" in our nation's schools.

The specific focus on changes in early childhood curricula has also seen widespread attention in the press. The Wall Street Journal suggests that you should "check out your neighborhood school. Reading, arithmetic and computers are fast replacing playtime in kindergarten. The four-to-six set spends more hours at desks, faces more rigorous tests and sits behind more computer screens that ever. It's even possible now to flunk kindergarten in places such as Minneapolis and Georgia" (Putka, 1988). Similar reports have been published in other

newspapers, ranging from Boston ("Yesterday's kindergarten program is now considered right for kindergarten" [Coons, 1987]), to Marin County ("Kindergarten isn't child's play anymore" [Cahil, 1988]), and from New York ("More than three million children are starting kindergarten this month, and for many it will be the first opportunity to fail" [Hechinger, 1988]) to California ("...possibly a fourth of the kindergarten population is not ready for an academic push..." [J. Fiske, 1988]). The notion that nearly one in every four children entering public kindergarten may fail or may not be prepared to benefit from the regular classroom program is startling. No evidence has been presented to support the large-scale policy of retention/extra-year placement. Indeed, the available data indicate without exception that retention is a policy that has negative effects on children (Bredekamp & Shepard, 1989; Charlesworth, 1989; Holmes & Matthews, 1984; Shepard & Smith, 1986, 1987; Smith & Shepard, 1987). Either we are witnessing a population shift of immense proportions, or we are experiencing a vast alteration in education policy—aided and abetted by the inappropriate use of standardized tests. This paper presents documentation that schools have changed, not children. Further, tests have contributed heavily to the shape and rationale for this change.

Why Schools Have Changed. One of the major sources of change in contemporary education is the pressure for accountability. The series of national reports that began with the National Commission on Excellence in Education's Nation at Risk (1983) called for standardized tests to be administered at all levels of schooling. The purpose of the tests was both to identify the need for remedial instruction and to "certify the student's credentials" (op. cit., p. 28). It is only a short step from this statement to the assumption that tests can be used to evaluate not only the student's learning but also the quality of the student's program and teacher.

In fact, an "accountability culture" has begun to emerge in our schools (Shepard & Smith, 1988). The pressure for teachers at one grade level to be held accountable, as measured by standardized tests, has resulted in an "academic trickle-down" process that has had a major impact on teachers in earlier grades (Cunningham, 1988). "As third grade teachers experience pressure for their children to perform well on standardized tests, they in turn put pressure on the

second and first grade teachers to prepare their children for the 'demands of the third grade curriculum'" (op. cit., pp. 24-25). Teachers' decisions about curriculum are thus influenced by a need for their students to perform well in the next grade level, a need that originates in part with the standards of accountability that are implied by standardized tests. In other words, teachers are very likely to shape their curriculum around a test's specific focus (see Darling-Hammond & Wise, 1985). This phenomenon, known as "measurement-driven instruction" (Madaus, 1988, p. 84), transforms testing programs, which should be the servants of educational programs, into masters of the educational process. This results in a narrowing of the curriculum, a concentration on those skills most amenable to testing, a constraint on the creativity and flexibility of teachers, and a demeaning of teachers' professional judgment.

Cross-National Studies. Research by Engel (1989) indicates that the controversy surrounding practices in early education are by no means limited to the United States. Engel looked at the following issues: age at school entry, measuring school readiness, ability grouping, kindergarten retention practices, and kindergarten curriculum in eight industrialized nations (the Soviet Union, Switzerland, West Germany, Sweden, England, New Zealand, Australia, and Japan). Although the study did not present a comprehensive picture of these practices in the eight nations, the results of the research provide an interesting perspective from which to view these practices in the United States.

The age at which children enter school ranges from four years (Britain and Australia) to seven years (Sweden). There are differences in performance between the older and younger children entering school in all the countries regardless of entry age, but these differences seem to disappear by about third grade. Interestingly, the entry age in the different nations has more to do with historical, political, and climatic reasons than with educational rationale. Six of these countries reported that ability grouping in the early grades does occur. Only New Zealand and England practice ability grouping, but in New Zealand the teachers form fluid groups based on observations of children, and in Britain the use of ability grouping has recently become less popular. None of the countries reported using standardized tests to group children.

Retention in kindergarten is viewed differently in the eight nations. For example, in both Japan and the Soviet Union grade promotion is automatic. The same is generally true for Australia. West Germany reports a retention rate between 5 and 10 percent, while in Switzerland the retention rate is quite high (in one canton it is 33%), but this policy is being reconsidered. Further, the curriculum of the kindergarten seems to be a source of debate in some of the nations. In England, the Soviet Union, and Japan, for example, there is a growing concern that kindergarten is becoming too academic.

The use of testing in kindergarten to measure school readiness, although not required by any of the eight governments, does take place, but the purpose varies by country. For example, the Soviet Union opposes the use of testing for any purpose but evaluating children who might be handicapped. The Swiss education laws are generally interpreted as stating that testing should be used to identify children with handicaps, but some interpret the law to mean that tests should also be used to indicate children's readiness for school, and tests are often used in this manner. In Sweden, the most widely used school readiness test, "Hostproven," which has little data supporting its validity, is used for diagnosis and curriculum planning at the beginning of the year. Switzerland, West Germany, and England all have a number of tests that are often used to assess school readiness. Entry into private and national public schools in Japan (1 percent of the elementary schools) requires that children take examinations. School readiness checklists, rather than standardized tests, are used in Japan and are also used in Australia and New Zealand. Although it is evident that tests are commonly used in these countries, general opposition to the use of standardized tests with very young children was reported in all but two of the countries studied.

Summary. At this point, three caveats must be raised about testing young children in American schools. First, the demographic changes in our society, particularly the changes in the composition of the workforce in the past generation that have resulted in more and more mothers of young children returning to work, have brought about an expansion in out-of-home care for young children. Children are entering kindergarten with two or more years of preschool or day

care experience and have had exposure to school-related tasks and routines. While this means that kindergarten-age children may "know" more and may even be somewhat more advanced developmentally, it does not follow that they are able to profit from modes and materials of instruction that are appropriate for children who are chronologically and developmentally a year or more older.

Second, schools may be pressured to adjust their curricula in order to meet standards of accountability, but these standards are typically driven by societal forces. Parents, school boards, legislatures, and governmental commissions all exert authority over the process and product of schooling. Ultimately, test manufacturers develop tests that reflect the priorities of these individuals and socio-political forces. But it remains the responsibility of professional educators to inform society about the best practices and most optimal objectives for children. Unfortunately, strong dissenting opinions have not stemmed the misuse of standardized tests—even when these voices have carried the imprimatur of the National Association for the Education of Young Children (1988), the National Association of Early Childhood Specialists in State Departments of Education (1987), and the National Association of State Boards of Education (1988). All three of these organizations have produced position statements calling for a more rational use of tests in early childhood. Related statements have been written by national organizations of school psychologists, elementary principals, and pediatricians.

Finally, not all tests are bad for kids. It is easy to be a "test basher." It is considerably more difficult to understand the complexities of psychometric research and the importance of selecting the right test for the right child at the right time. Specifically, reliable and valid developmental screening tests, when administered to individual children by trained testers, can be used to identify children who are at high-risk for school failure (see Meisels, 1988, 1989a, 1989b, 1989c; Meisels & Wasik, 1990). Children so identified would move on to a more-comprehensive diagnostic process to determine conclusively the nature of their problems and subsequently, to obtain appropriate interventions. In other words, the problem is not tests per se, but the appropriate and inappropriate use of tests in specific situations by specific individuals.

C. High-Stakes Testing.

The kinds of tests that have created a crisis in public early childhood education are readiness or achievement tests that are used for classification, retention, or promotion. Tests used in this manner can be described as high-stakes tests: "those whose results are seen—rightly or wrongly—by students, teachers, administrators, parents, or the general pubic, as being used to make important decisions that immediately and directly affect them" (Madaus, 1988, p. 87). The high-stakes decisions that flow from such tests concern retention, promotion, placement in pre-kindergarten or pre-first grade programs, evaluation and rewards for teachers or administrators, and allocation of resources to specific schools or school districts. Three specific characteristics of high-stakes tests that have been analyzed previously (Meisels, 1989a) will be described.

Perceptions. High-stakes tests often achieve a life of their own, in which the tests' original purposes are blurred and their results begin to assume greater importance than ever imagined by those who proposed them. The SATs are the best example of this phenomenon, in which a test that was intended to provide supplementary information to assist in decisions regarding college admission has not only become an absolute criterion for admissions decisions in many cases, but has become a barometer of the entire nation's educational progress. Similarly, when actions are taken that have an impact on the results of high-stakes testing, e.g., instituting a preparatory course designed to boost SAT scores, it is assumed that the underlying skills and abilities measured by the test have been changed, rather than the test-taking skills improved. As Madaus (1988) puts it, "People fail to distinguish between the skill or trait itself and a secondary, fallible indicator or sign of them" (p. 90).

Instruction. The corollary to this is that high-stakes tests have a major influence on teachers' behavior and on their instructional decisions. It is virtually a maxim of American educational research and practice that a teacher's perceptions of a child can be heavily influenced by the child's race, sex, socio-economic status, and by quantitative measures that purport to assess the child's potential (see Brophy, 1983). If we manipulate any of these variables we are likely to alter teachers' attitudes and behavior towards their pupils. In a similar manner,

teachers' instructional decisions can be affected by the tests they use. If teachers know that their pupils will be tested on certain skills or certain subject areas, and if the results of the examinations are to be made public, it is very likely that the teacher's curriculum will reflect these test-specific characteristics. This is an example of measurement-driven instruction, a concept mentioned earlier. High-stakes achievement tests invariably narrow instruction and learning, focusing the curriculum on the content that will be included on a test (see Koretz, 1988).

Decision-Making. Another characteristic of high-stakes testing is that these tests transfer control over the curriculum to the agency that sets or controls the exam (Madaus, 1988). Given the previous statements about the potential effects of using high-stakes tests, it is clear that test developers have a powerful role in shaping instructional and other educational decisions. In high schools and colleges one can assume a certain minimal consensus about content. However, in early childhood, no such consensus exists. For example, there are many ways to learn to read. A test that focuses on children's knowledge of sight words may overlook their ability to decode, use phonic skills, or engage in activities associated with emergent literacy programs (see Teale & Sulzby, 1986). Yet, if a school district adopts a test that reflects a particular approach to reading, teachers may feel enjoined to teach that approach. Hence, educational decision-making is removed from the arena of teacher/child interaction and is supplanted by the instructional approach implicit in the high-stakes test that has been selected for the school or the district.

This situation raises the concerns of many parents, professionals, and policy makers. Yet the tide of expanded testing keeps rising, and the implications of making educational decisions based on many of these tests becomes increasingly alarming.

Examples of High-Stakes Testing in Kindergarten. Many examples of tests that have achieved high-stakes status in early childhood programs can be presented. Three specific tests and a state testing program will be reviewed briefly in order to illustrate the impact of high-stakes testing in the early childhood years.

The Gesell School Readiness Tests

Madaus (1988) suggests that the power of high-stakes testing is "a perceptual phenomenon: if students, teachers, or administrators believe that the results of an examination are important, it matters very little whether this is really true or false—the effect is produced by what individuals perceive to be the case" (p. 88). This principle is clearly embodied in the widespread adoption of the Gesell School Readiness Test (Haines, Ames, & Gillespie, 1980). The problems with the Gesell are extensive, and have been described at length elsewhere (Bear & Modlin, 1987; Bradley, 1985; Kaufman, 1985; May & Campbell, 1981; May & Welch, 1984a, 1984b; Meisels, 1987; Naglieri, 1985; Shepard & Smith, 1985, 1986, 1987; Smith & Shepard, 1987). Its principal fault lies in the discrepancies between its stated purposes and the empirical evidence available to support those statements. Clearly, the Gesell is a high-stakes test: it promises to identify children who are at high-risk for school failure, and it asserts that it can be used to determine when children should begin school, which children should be promoted, and which should be retained in grade.

Unfortunately, there are no data to support these assertions. In one study which paradoxically claims to validate the Gesellian concept of developmental age, Wood, Powell, and Knight (1984) found that more than half of those kindergarten-age children who were considered "ready" by the Gesell did not have successful kindergarten experiences, as reported by their classroom teachers. A second study by May and Welch (1984b) also revealed major problems with the Gesell's accuracy, and found no support for the effectiveness of an extra-year program based on Gesell recommendations. Other studies with similar results are reviewed in the publications noted above. In short, these studies demonstrate that the claims of the Gesell theorists cannot be supported by empirical data.

Yet, the tests continue to be widely used—based, perhaps, on the unfounded perception that they are efficacious and because they provide a means for teachers to cope with the process of "academic trickle down," the inappropriate curriculum demands that they must endure and implement. In other words, if, as the Gesell theorists claim, their test measures "developmental

age," which is maturationally driven and genetically derived, a child who cannot cope with an academically-oriented school curriculum does not necessarily represent a failure on the part of the child, teacher, or parent. Rather, the child is simply not "ready," and no amount of instruction, intervention, or effort can be expected to have an effect. But this assumes that readiness is an absolute concept, not a relative one. Bruner (1966) notes that the idea of readiness is a "mischievous half-truth. It is a half-truth largely because it turns out that one teaches readiness or provides opportunities for its nurture, one does not simply wait for it. Readiness, in these terms, consists of mastery of those simpler skills that permit one to reach higher skills" (p. 29). When the Gesell tests are used to define readiness, not only has the concept of readiness been reified and misrepresented, the stakes have become very high indeed.

Use of Readiness Tests for Instructional Decisions.

Consistent with Bruner's perspective, the purpose of readiness tests is to evaluate a child's relative preparedness to profit from a specific curriculum (see Meisels, 1986, 1989). Most readiness tests are described as criterion-referenced instruments—those in which a particular child's score is indicative of a specific level of concurrent performance mastery. In contrast, norm-referenced tests are interpreted on the basis of a child's standing in relationship to a larger population or group of children (see Angoff & Anderson, 1967; Barnes, 1982). Predictions about future performance can be made based on this standing. Thus, the basic purpose of criterion-referenced tests is to measure current achievement, not to predict future performance.

It follows, therefore, that the use of criterion-referenced readiness tests for high-stakes purposes of classification, retention, and promotion is unjustified. The Brigance K & 1 Screen (Brigance, 1982) exemplifies this problem. The Brigance is a brief inventory designed to provide a general picture of a young child's language development, motor ability, number skills, body awareness, and auditory and visual discrimination. Based on its content and its criterion reference the Brigance is a readiness test, rather than a developmental screening instrument. Nevertheless, the Brigance is in very wide use nationally to make predictions, that is, to "rank or group children who are high, average, or lower than their local reference group in order to

contribute to readiness decisions, to make placement decisions, and to serve as an indicator for more comprehensive evaluation or referral for special services" (Boehm, 1985, p. 224). In order to fulfill these purposes, it is necessary that the test be norm-referenced and that it be accurate, so that high-stakes decisions will not be based on misleading data.

However, no reliability, validity, or standardization data are available for the Brigance. The test consists of a number of characteristic traits, skills, and behaviors that children at different ages demonstrate. To assume that this unstandardized collection of criterion-referenced items gives a definitive picture of a child's future ability is highly questionable. Furthermore, high-stakes testing carries high-stakes consequences for the tester as well as the child. As one review cautions, the lack of standardization data for the Brigance suggests that "any school system that formally and systematically uses the Brigance inventories without going through a local validation effort is placing itself at risk legally" (Robinson & Kovacevich, 1984).

Given this background the use of the Brigance for instructional decision-making is also questionable. Indeed, most achievement/readiness tests are of limited relevance to teachers because they assess a restricted range of instructional objectives, they omit major adaptive and socio-emotional behavior, or they are perceived as doing little more than confirming what the teacher knew about the child already (Durkin, 1987; Kelleghan, Madaus, & Airasian, 1982; Salmon-Cox, 1981). The missing ingredient is the match between the test and the teacher's curriculum. To the extent that the test reflects the teacher's approach and instructional goals, it is likely that it will have a positive impact on educational decision-making. When readiness tests are used for low-stakes internal testing programs they are often not perceived as particularly efficacious because of this lack of fit with the teacher's goals, and they usually do not have a major impact on instruction. Yet when the same tests are transformed by administrative decree into high-stakes tests, they can and do influence instruction, though clearly not for the right reasons.

The Georgia Experience.

The final characteristic of high-stakes testing to be discussed concerns the subtle transfer of control over the curriculum to the test developer. Nowhere is this abrogation of instructional authority better exemplified than in the testing program implemented by the state of Georgia in 1988. In 1986 the state passed a bill known as the Quality Basic Education (QBE) Act. This bill required all children seeking to enter first and fourth grades to pass a test that would demonstrate their academic readiness. Students who did not pass such tests and, in kindergarten, whose teachers confirmed the results of the readiness assessment, would be required to repeat kindergarten or third grade. Because of the national outcry concerning this program, Georgia recently announced that it will institute a revised testing program next year. Nevertheless, there are many lessons to be learned from the original Georgia plan, and they will be reviewed below.

The test selected for first-grade entry by the Georgia Department of Education is the California Achievement Test (CAT), level 10 (CTB/McGraw-Hill, 1988). In the "Georgia Edition" of the CAT, however, only 64 of the 146 items (44 percent) are administered. The stated purpose of the Georgia CAT is to measure achievement in the basic skills and to provide specific information about students' instructional needs. The manual states that the CAT items "may be used to establish reference points for beginning instruction in kindergarten and to predict first grade reading achievement" (CTB/McGraw-Hill, 1988, p. 1). Thus, the Georgia CAT is a high-stakes test: it is designed to render decisions about student classification, retention, and promotion; it is intended to guide instructional decisions; and it is perceived as carrying out the state's mandate to establish quality education programs. Unfortunately, the test and the testing program fall far short of achieving these goals. An analysis of nine of these shortcomings demonstrates clearly how high-stakes early childhood achievement testing can have potentially deleterious effects on a public system of education.

First, the test was modified without any empirical validation, although it is a psychometric axiom that subsets of items do not share the psychometric properties of the core test from which the items were drawn (APA/AERA/NCME, 1985). The entire test was piloted in Georgia (the

complete test takes nearly 3 hours to administer), but no specific validity data were reported about the subsample of items that were selected.

Second, the Georgia CAT represents a very narrow view of learning, as only the three subtests of visual recognition, sound recognition, and mathematical concepts are included, constituting a limited focus on literacy and numeracy. Missing is any assessment of the child's attention, motivation, expressive language, motor development, use of materials, rate of learning, preferred modality, etc.

Third, the enterprise of whole-group standardized testing in high-stakes testing is questionable. Wodtke, Harper, Schommer, and Brunelli (1985) conducted a study of teachers' group testing practices in eight kindergartens. Their findings revealed wide variation in testing conditions, many departures from standardized testing procedures, and extensive variation in children's behavior. This study highlighted the variability, lack of objectivity, and the dependence on context of standardized testing. The Georgia CAT, which is a whole-group administered test, is subject to the same type of variability and limited reliability.

Fourth, the decision mechanism of the test is unstandardized. As originally conceived, children are administered the CAT and assessed by their teachers. If any discrepancy exists between the standardized assessment and the non-standardized teacher report, then the child would be administered another readiness test. But all three of the assessment procedures are of unknown reliability and validity. Thus, an unstandardized test is to be accompanied by an idiosyncratic, non-systematic teacher report form, which may be followed by testing with another non-standardized instrument that may be measuring different phenomena altogether!

A fifth problem concerns the establishment of cut-off points to indicate failure. Initial results indicate that eight percent of the children who took the CAT in 1988 failed to score above the Georgia cut-off, that is, the tenth percentile. In some districts the failure ratio was as low as one percent; in others it was as high as 26 percent (Cunningham, 1988). No data are available concerning the racial, ethnic, geographic, and socio-economic composition of this group of children. It is possible that poor and minority children are overrepresented among these

"failures" and that the lack of cultural sensitivity of the test may have contributed to this problem. In any event, if as many as eight percent of the children were unable to perform above the tenth percentile on this test, it is clear that the school districts had not previously identified those children who were at high risk for school failure. Use of a validated developmental screening test at the outset of kindergarten (see, Meisels, 1985) could have resulted in most of these children being identified before they experienced a year of kindergarten failure.

The sixth concern relates to the consequence for failing the CAT—retention in grade. The evidence concerning kindergarten retention does not support its use for improving academic achievement (Plummer, Lineberger, & Graziano, 1986; Shepard & Smith, 1987, 1988). Indeed, it is likely that retention under these circumstances may result in lowered self-esteem and rejection by the child's peer group, issues which overshadow any short-term academic gains.

Seventh, the Georgia law has the potential for creating a highly stratified, homogeneous group of children who are retained in grade. One might ask why these children should not be mainstreamed. One must also be concerned about the potential long-term consequences of being a year or more older for grade than one's peers. According to a recent report of one large city school-sponsored task force, age/grade status is the single most sensitive indicator of dropout potential in urban school districts (Detroit Dropout Prevention Collaborative, 1987). Of those students who were at least one year overage in ninth grade, more than 45 percent dropped out of school by twelfth grade.

Eighth, the state has imposed the QBE, but it has not provided financial resources to support its implementation. No new funds are available to school districts for remedial programs, new materials, or hiring staff in order to reduce class size and improve teacher-child ratios.

Finally, the Georgia plan abridges parental and children's rights that were secured nationally in the 1970s. The Georgia program does not grant parents the right of appeal or of due process; it permits placement decisions to be made on an arbitrary and capricious basis by classifying children with a non-scientific and invalid test; and it flies in the face of provisions for

the least restrictive environment, parent participation, and the use of validated tests from multiple sources and multiple disciplines that are fundamental to Public Law 94-142 (Gartner & Lipsky, 1987; Heller, Holtzman, & Messick, 1982).

The action of the Georgia legislature in promulgating the QBE should serve as a warning to parents, professionals, and lawmakers throughout the nation. The Georgia plan for kindergarten testing and retention defames the importance and value of accurate educational measurement. It is the reductio ad absurdum of high-stakes testing in which an entire state (and so far the only state) has transferred control over its early education program to a single group-administered, paper and pencil test. Teachers have begun to alter their curriculum and their teaching styles so that children will have a better chance to do well on the test. Administrators and teachers in local school districts have been told that their performance will be evaluated by the gains made by their students on the CAT in succeeding years. Private firms have begun to offer preparatory classes to kindergarteners (called "CAT Academies") so that they will pass the test, and national companies have begun to market kindergarten beginning test-taking skills programs.

Although Georgia is the only state to have instituted a policy whereby every child must pass a readiness test or else repeat kindergarten, a recent study commissioned by the National Academy of Sciences and the National Association of State Boards of Education documented the existence of kindergarten testing policies in more than 30 states (Gnezda & Bolig, 1988). The study also noted that 43 states reported that some districts use academic readiness tests prior to first grade, and 40 states reported that their local districts sponsored developmental kindergarten or transitional first grades in some of their schools.

However, as the authors of the report state, "Early in the data collection phase it became clear that in all states the majority of testing decisions are made locally with minimal, if any, input from the state level" (Gnezda & Bolig, 1988, p. 2). Thus, most of the essential data needed to analyze the impact of testing on young children can only be obtained from local education agencies (LEAs).

Despite the absence of such data—or perhaps because of it—much of what we know about the extent of early childhood testing, tracking, and retention has emerged from reports in local and national newspapers and national news weeklies, as noted earlier (see, for example, Coons, 1987; Carmody, 1989; Fiske, E.B., 1988; Fiske, J. 1988; Hechinger, 1988; Ordovensky, 1989; Putka, 1988). The absence of systematic data about testing, tracking, and retentions greatly impairs the development of policy alternatives. Although a strong suspicion exists that the way in which standardized tests are being used is having a negative impact on children, schools, and teachers—and particularly on minority children—the extent, range, and intensity of these problems are unknown.

D. Tests

Just as there is confusion about the extent and impact of early childhood testing, so is there confusion about what is meant by "testing." As noted earlier, more testing is taking place in early childhood and kindergarten than ever before. Further, more young children are being classified and placed in extra-year early childhood programs because of the inappropriate use of tests and this is probably happening disproportionately with minority students.

As tests assume a greater role in the early childhood educational process there appears to be a tendency to rely on tests alone to make educational decisions. Yet, whenever possible, test information should be supplemented with data derived from parents, teachers, other professionals, and first-hand observations (Meisels & Provence, 1989). The task of keeping testing in perspective includes a recognition of the following (see Meisels, 1989).

1. Tests do not have magical powers.

A test does not in itself have power, nor does it automatically convey power to its users. Tests are only powerful if we transfer to them our control over decisions regarding what is to be taught, what is to be learned, who is to be promoted, or who is to be retained. Tests can assist us in making these decisions. But they need not be the masters of the educational process; they should facilitate that process (Meisels, 1989a).

2. There are various types of tests; testing is not a monolith.

The principal types of tests that are useful to early childhood educators are developmental screening tests and readiness/achievement tests (Meisels, 1986, 1987a, 1987b, 1988, 1989a). These tests differ in very significant ways from one another and should never be used interchangeably. Specifically, readiness tests should never be used to predict a child's future potential. Only a valid and reliable developmental screening test can serve this purpose.

3. It is essential that tests only be used for their intended purposes.

Both developmental screening tests and readiness/achievement tests have a role to play in early childhood programs. But they serve different purposes. Screening tests help select children who are likely to be in need of special services because of a learning problem or handicapping condition. Only developmental screening tests that are reliable and valid should be used. Readiness/achievement tests can determine a child's relative preparedness to participate in a particular classroom program, or can document a child's acquisition of skills and knowledge (Meisels, 1989b; Meisels & Provence, 1989).

4. Tests should not be used to make high-stakes decisions in early childhood programs.

High-stakes tests are those that are directly linked to decisions regarding promotion or retention, that are used for evaluating or rewarding teachers or administrators, that affect the allocation of resources to programs, and that result in changes in the curriculum (Madaus, 1988; Meisels, 1989c). None of these decisions should be controlled solely by tests in early childhood. Rather, if such decisions are undertaken, tests should only provide supplementary information to help the teacher, parent, and other specialists arrive at the best possible decision for each child.

5. Instructional decisions and documentation of accountability should be based on teacher-derived information, rather than on test data.

Early childhood programs should focus on the teacher's contributions to instructional decision making and accountability rather than relying on tests to perform these functions. "Measurement-driven instruction" and "test-based accountability" distort both the test's importance and the teacher's role. For purposes of instruction and accountability more emphasis should be placed on enhancing the teacher's "kid watching" abilities, and systematic means of recording teacher's observations of children need to be devised (Cunningham, 1988; Meisels, 1989c; Shepard & Smith, 1986). All of these concerns become even more important when the situation of children from minority backgrounds is explored.

E. How schools, teachers, and tests are failing minority children

The use of tests with minority students is controversial, reflecting the belief that tests do not measure what they are purported to measure when used with children from cultural, ethnic or social class backgrounds that are not mainstream. California has, in fact, outlawed the use of standardized individual tests of intelligence with black students for any purpose (Dent, Mendocal, Pierce, & West, 1987). This decision was the result of a class action suit brought by parents on behalf of their children who had been misclassified in educable mentally retarded (EMR) special education classes in San Francisco. Consequently, school districts in California must now devise alternative methods for determining the educational needs of black students.

One criticism that is often levied against tests is their potential communication and language bias (Taylor & Lee, 1987). For example, it is suggested that different portions of the population may have different cognitive styles. Tests that are constructed to reflect a particular cognitive style would be measuring different ways of knowing and problem solving, rather than assessing ability, which, of course, is what the test purports to measure. In order to guarantee culturally fair standardized tests, it would be necessary, Taylor and Lee (1987) suggest, to accept a variety of response types and to have a variety of tasks to elicit a single response.

Teacher perceptions are another bias that negatively affects achievement in minority children. Poor and minority students are less likely to be placed in programs for gifted and talented students, they are disproportionately enrolled in special education, they are overrepresented in vocational education programs, and they are underrepresented in academic programs (Chunn, 1988). In general, teachers expect black children to do less well than white students, nonstandard English speaking students to do less well than standard English speaking students, and low income students to do less well than middle income students. These perceptions result in children being tracked into low ability groups early in their education. These low ability groups have been criticized from several vantage points, for example, they tend to be much more disruptive as learning environments than high ability groups. Research has indicated that for the majority of children, heterogeneous groups are a preferable way to group students in school (Chunn, 1988). The effect these teacher perceptions have on students' scores on standardized tests is, of course, an important question.

Pallas, Entwisle, Alexander, and Cadigan (1987) examined the variables associated with children in first grade who do exceptionally well according to their scores on the California Achievement Test verbal section. In general, they found that the teacher's rating of the child's personal maturity and the teacher's assessment of positive school climate were highly associated with children doing exceptionally well in first grade. In other words, how a teacher perceives a child's maturity level and how she perceives the environment she is teaching in is more influential of scores on standardized tests in first grade than the child's marks, or background variables such as how often a child is read to at home.

A question that arises concerns the qualities in a teacher that might result in her having negative perceptions of black students' ability level and a negative impact on these students' success on standardized tests. Alexander, Entwisle, & Thomson (1987) compared teachers' family of origin SES with the perceptions they had about the school they worked in and their students. In general, it appears that teacher SES, not race, interacts with students' race, not their SES. In other words, high SES teachers, regardless of their race, seem to have lower

expectations for black students. The authors conclude, "They [high SES teachers] perceive such [black] youngsters are relatively lacking in the qualities of personal maturity that make for a 'good student,' hold lower expectations of them, and evaluate the school climate much less favorably when working with such students. As a result, blacks who begin first grade with test scores very similar to their white age-mates have fallen noticeably behind by year's end" (p. 679).

A final variable that seems to influence the performance of minority children on standardized achievement tests is examiner familiarity. In a meta-analysis of 14 studies looking at examiner familiarity in primarily preschool and elementary school testing, Fuchs and Fuchs (1989) concluded that the use of familiar examiners has a significant effect on the scores of minority youths, but did not seem to have an impact on white children's performance. They claim that on a typical standardized IQ test, use of a familiar examiner would raise a minority child's score from 100 to 111, while it would have virtually no effect on a white child's score. The results of this study prompted the authors to conclude that "comparing minority student's suboptimal performance with unfamiliar examiners to the more maximal performance of largely Caucasian normative populations could result in spuriously low and improperly restrictive educational placements of minority children" (p. 307).

These observations justify a closer look at the interaction between race and achievement, with particular attention being paid to the role of standardized tests. Alexander and Entwisle (1988), in their longitudinal study of the first two years of schooling, explored the effects of schooling on academic achievement. One of the most salient aspects of their findings is the change in achievement and expectancies that occurred for black children and their families from the beginning of first grade to the beginning of second grade.

Over 800 first graders, with their parents and teachers, were randomly chosen to participate in this study of achievement in the first two years of school. These 800 students were administered the California Achievement Test (CAT) during the fall of their first grade year. There were no significant differences in performance on these initial CAT scores due to race or

gender among these first graders, although small effects for parents' beliefs about their child's ability and for parents' expectancy about school success were observed.

The similarities between white and black first graders were replaced with important and significant differences by the first marking period (within the first 3 months) of first grade. For black children the CAT scores obtained as school started were not predictive of the first marks given by the teachers. Yet these first-quarter marks were strong predictors of black children's second year CAT scores. Specifically, these first quarter grades were twice as predictive for black students of CAT gains in math scores at the beginning of second grade. These same first quarter marks predicted retention at the end of first grade, although CAT scores were not predictive of failure (Cadigan, Entwisle, Alexander, & Pallas, 1988). In fact, retention at the end of first grade was predicted by first quarter reading marks and by the questionnaire administered to teachers about their perceptions of these students' abilities while they were in kindergarten. These teacher-based judgments resulted in 16 percent of the students being retained by the end of their first-grade year, independent of scores on the CAT.

When looking for student behaviors that put children at risk for academic success, Cooper and Farran (1988) found that teacher perceptions and rating of students have high stability throughout the year and that one of the strongest predictors for high-risk status among kindergarten children is "maleness." From the longitudinal work of Alexander, Entwisle, Cadigan, and Pallas (1987) it appears that "blackness" or "poorness" are other risk factors for school success. They found that teachers' values affect their evaluation of student performance, and these values interact with the SES of the children. High SES teachers rated lower SES students more negatively than other students. The researchers summarize their findings by saying, "Teachers identify some students as losers from the start" (p. 76). Teacher effects are not associated with year-end achievement test scores in this large sample but with the grades given by the teachers.

The power of others' perceptions on student performance is dramatically demonstrated in the data for math performance between boys and girls. Alexander et al. (1987) found that white

parents' expectancies for boys and girls for math achievement was the same at the beginning of first grade. There were no differences found, on average, between boys and girls on achievement tests or marks given by teachers. At the beginning of second grade, white parents reported higher expectancies for their boys than for their girls on math achievement. At the end of second grade, boys' math scores had surpassed the girls' scores on math. It is difficult to believe that anything other than expectancy affected this change in math scores for girls in the period from the end of first grade to the end of second grade.

Another important change which occurred between the fall of first grade the fall of second grade was the focus of the black parents away from their children's abilities, to a "preoccupation" with retention status (p. 102). Parsons, Adler, and Kaczala (1982) found that parent perceptions of children's abilities have more powerful effects on student achievement than childrens' "actual" abilities. Entwisle and Hayduk (1982) found that when parents believe their children are smarter than other children, these children do better than other children. If, despite achievement scores, children are retained by the end of first grade, it seems likely that parents would be concerned about this status, and doubt their own beliefs about their children's abilities. It follows that if parents' positive beliefs have positive academic outcomes for children, parents' negative beliefs about ability would have negative academic outcomes.

Teacher perception of childrens' abilities seems to have an undeniable effect on whether or not children can be successful in the school environment. In a longitudinal study of black children in a segregated urban school, Rist (1970) found that kindergarten teachers made evaluations of students' expected abilities based on appearance, language style, and SES characteristics of the families. Without any indication of these new students' academic ability (as measured by preschool attendance, screening tests, etc.) the teacher that Rist studied placed children in one of three groups, based on her perception of whether or not they were "fast learners." All of the "fast learners" were perceived by the teacher as clean and neatly dressed, spoke standard English, interacted verbally with the teacher, and had families which were intact and not on welfare. The hierarchical placements made by the kindergarten teacher remained

intact throughout the year, with no child moving from one group to another. Moreover, when these children were in the second grade, the second grade teacher grouped and seated them in the same hierarchical manner. The only changes in placement that occurred from the first week of kindergarten until the end of second grade took place in December of the second grade year. Two students seated at the "fast learner" table were moved down one table because "they kept their table and floor messy." Two students from the second group moved up to the "fast learner" table because they kept their table and floor neat.

This study shows the futility of making important judgments about students' academic abilities based on grades and tests when we see how easily teachers make judgments about academic ability based on how children are dressed and the SES of their families in the first week of kindergarten. Studies by Alexander and Entwisle, Rist and many others demonstrate the pervasiveness of these practices. When children are placed in learning groups based on these non-academic and unchangeable attributes, there is little hope of their ever moving out of the "low" group into another group of learners.

When we examine how teachers treat the children in the "fast learners" groups with regard to amount of time spent on engaging children in the teaching/learning process, giving support and help for academic work being done by the children, and providing opportunities to demonstrate what they know by being called on and asked to participate in group projects, it is clear why "fast learners" continue to succeed in school while the gap in the other childrens' achievement continues to widen with each year in school.

The complex psychological web that is generated beginning in the first week of kindergarten and that prescribes success and failure in school life cannot be blamed solely on teachers, parents, or tests. But teachers and other school personnel must cease consigning groups of children to a poor education by making judgments of the children based on anything that denies them the chance to see themselves as able and equal participants in the teaching/learning process. Tests, teacher perceptions, and retention are too powerful to be used as weapons against children.

F. Why Test Young Children?

Given this background concerning the misuse of tests and other associated problems, particularly for children from minority backgrounds, one can question the value of testing young children at all—especially if the results are used for high-stakes purposes. One such purpose is that of retention. Shepard and Smith (1986) provide a thorough analysis of the relationship between readiness testing and kindergarten retention policies. In addition, they examine the research literature about the "problem of being youngest." They conclude that children should not be retained in kindergarten or placed in a two-year pre-kindergarten or readiness program based on the use of readiness tests alone. These tests are insufficiently accurate to be used for screening and placement. Moreover, they cite research that shows that when such tests are used to assign children to extra-year programs, these programs contribute to children's lower self-esteem, rather than their higher achievement. Finally, they note that the rationale for such programs—to give younger children time to mature—is not supported by research. "The disadvantage of the youngest first graders is small...the youngest problem will disappear by third grade unless it is cast in stone by a learning disability label or grade retention" (op. cit., p. 83). Salzer (1986) also comments on the limited accuracy of readiness tests and the potential costs of labeling children. His recommendation is to focus instructional attention on children's strengths, rather than their weaknesses. He sees the "test-teach-test" model in use in many school districts as inherently limited and short-sighted.

These papers, and others like them (e.g., Bredekamp & Shepard, 1989; Charlesworth, 1989; Cunningham, 1988), make critically important points about testing young children—points that should be considered by everyone who establishes policies for young children. Chief among these points is that school readiness tests can not be used appropriately for prediction and class placement. The data obtained by means of such tests—e.g., the Metropolitan Readiness Tests (Nurss & McGauvran, 1976), the Gesell School Readiness Test (Ilg & Ames, 1972), and the Cognitive Skills Assessment Battery (Boehm & Slater, 1978)—are intended to describe a child's current level of skill achievement or pre-academic preparedness. These entry level skills are not

strongly associated with those outcomes that are measured by tests, grades, or retention practices (see Meisels, 1986). If one's goal is to predict quickly whether a child might have difficulty succeeding in school, or could profit from a specialized educational placement, then a different kind of test must be used: one with predictive validity, developmental content, and normative standardization. Tests that have these properties are known as developmental screening tests. Examples include the Early Screening Inventory (Meisels & Wiske, 1983), the McCarthy Screening Test (McCarthy, 1978), and the Minneapolis Preschool Screening Inventory (Lichtenstein, 1980).

Thus, the answer to the question, why test young children? depends on the goals of the individuals who select and administer the tests. Different goals call for different kinds of tests, and some of the most common abuses of testing are attributable to the use of tests in situations for which they were not designed.

What Kinds of Tests Should We Use? Developmental screening tests and school readiness tests represent the two most widely used kinds of tests for pre-kindergarten and kindergarten-age children. Neither test should ever be used to label children or assign them to diagnostic categories. But beyond this similarity these two types of tests differ from each other in purpose, content, standardization procedures, and psychometric properties. Developmental screening tests are used to identify children potentially in need of special education services. Readiness tests focus on a child's relative preparedness for benefiting from a specific pre-academic program or curriculum. Developmental screening tests reflect a child's ability or potential to acquire skills, while readiness tests identify a child's current skill achievement, performance, and level of general knowledge. Screening tests are norm-referenced and must have excellent reliability and predictive validity. In contrast, readiness tests are typically criterion-referenced, and have reliability, but usually only construct validity (see Meisels [1984; 1989] for explanations of these terms).

These differences between the two kinds of tests underlie the differences in their use. Developmental screening tests are intended to predict which children will be high-risk or

handicapped—although only screening tests with well-established validity can accomplish this goal. Readiness tests should not be used for prediction or placement. They inform us about a child's current status, but give us little information about a child's potential to move to another level of skill accomplishment. The differences between these two types of tests—and their similarities—are highlighted in the discussion below.

Norm-Referenced and Criterion-Referenced Instruments. Probably the most widely used type of test for measuring children's progress is the norm-referenced test. The major characteristic of norm-referenced tests is that their scores are interpreted on the basis of the standing of an individual child with respect to some larger population or group of children (Barnes, 1982). In a norm-referenced test, the average performance of the subjects in a standardization sample becomes the basic reference point or norm against which future individual scores or performances will be compared (Angoff & Anderson, 1967). When using norm-referenced tests, it is essential to have information about the original standardization sample in order to interpret the data obtained from that test. Thus, one of the major issues confronting early childhood educators who choose to use norm-referenced tests to assess a child's progress is the problem of finding a norm-referenced test with an appropriate reference population (Hamilton & Swan, 1981; Meisels, 1987b).

Programs that utilize norm-referenced measures as indicators of child progress may thus be making unexamined assumptions about the meaning of these findings. Even if one uses these instruments within a homogeneous population, untested assumptions about the meaning of the results remain. Chief among these unexamined assumptions is that intraindividual progress among children will follow a linear pattern, such as that suggested in the pattern of items included on most normative scales. When one uses norm-referenced tests with a heterogeneous population, the problems of intraindividual difference become more exaggerated. Use of such tests in this situation would seem to imply an assumption that program effects are similar for all subjects. But, as noted above, it is likely that program effects differ as a consequence of subjects' experience, as well as other important demographic factors.

An alternative to norm-referenced testing is the use of criterion-referenced instruments. The score that an individual child obtains on a criterion-referenced test is indicative of a specific level of performance, or a specific degree of mastery. There are two types of criterion measures: domain-referenced and objective-referenced tests (Barnes, 1982). In a domain-referenced measure, the score a child obtains indicates the proportion of a specific domain or subject area that the child has mastered. Typical examples of domain-referenced instruments include tests of spelling, arithmetic, reading, and other domains with clearly defined bodies of knowledge (Hamilton & Swan, 1981). A score on a domain-referenced test has an absolute meaning, indicating the extent of mastery of a specific area demonstrated by an individual. In contrast, objective-referenced tests refer to specific objectives, drawn from a larger set of possible items, that are to be achieved by the child (Barnes, 1982). Mastery on an objective-referenced test is defined either by a perfect score on these selected items, or by achieving a predetermined proportion of successes, e.g., four successes on five trials for each of the selected items.

But it is misleading to believe that criterion-referenced tests do not also require rigorous standardization, as do norm-referenced tests. Because in utilizing a criterion-referenced behavioral-objectives test in which one accepts, for example, a 90 percent criterion as mastery for some particular skills, one is implicitly using normative data as the basis for establishing a standard or criterion of performance. In other words, the use of such a criterion would imply that mastery is defined as that level which 90 percent of a normative group has mastered or passed. Otherwise, there would be no rational way to establish reasonable goals. Thus, normative- and criterion-referenced tests are not mutually exclusive. If a criterion-referenced test does not have well-established norms, it may result in the establishment of unreasonable expectations for the users of that test. In many respects therefore, criterion-referenced tests can be seen as a special instance, or category, of norm-referenced tests, even though the two types of tests are used for different purposes and yield different information.

There are two areas of critical difference between norm- and criterion-referenced tests. One such difference is the issue of variability (see Popham & Husek, 1969). Scores on norm-

referenced tests are meaningful only in terms of the relative position of a score when compared with other scores. Thus, the more variability in those scores, the better the test. But variability is irrelevant to criterion-referenced tests because the meaning of the score is independent of comparisons with other scores. Rather, the meaning of criterion-referenced scores is directly related to the relation between the items and the criterion. Thus, one can have a useful and worthwhile criterion-referenced test with very low variability of scores; either individuals master a particular objective or task, or they do not.

Another major difference between the two types of tests lies in their relation to prediction. The basic purpose of criterion-referenced tests is evaluation and measurement of current performance, not prediction of future performance. Because such tests are intended to describe what the child is capable of doing right now when compared to a specific criterion, predictive validity has little meaning for criterion-referenced tests. Although criterion-referenced scores may be correlated to some other future event or circumstance, this correlation is coincidental to the major purpose of the measure (see Barnes, 1982). The criterion of the criterion-referenced measure is in the here and now, not in future performance, as is the case with norm-referenced tests.

The use of readiness tests as predictors of school success is beset with problems: children who can do well in regular classes are misidentified as "slow" or "developmentally immature", while children who could profit most from an individuality or special education program may be missed altogether (Meisels, 1987a, 1989c). Furthermore, issues concerning chronological vs. developmental age have become almost hopelessly entangled by some advocates of readiness testing because all too often younger children who score low on readiness tests are labeled "developmentally immature" (the Gesell is a good example of this), and will be placed in "developmental readiness" classes. However, readiness test content is not in fact developmental, but is more closely related to the impact of direct instruction on skill acquisition. Thus, children who may simply need an individuality program of skill development are being erroneously labeled and/or retained in grade.

In short, two kinds of tests can be of value to educators working with young children, developmental screening tests and school readiness tests. But one cannot be substituted for the other. Screening tests provide a brief assessment of the developmental abilities highly associated with children's future school success. Readiness tests are concerned with which curriculum-related skills a child has already acquired. If a school administrator or teacher wants both kinds of information, then both kinds of tests should be administered.

Which Tests Should We Adopt? After making a decision about what kind of test to administer one of the next questions concerns which test to adopt. Descriptions of screening and readiness tests are available from many sources (e.g., Barnes, 1982; Lichtenstein & Ireton, 1984; Meisels, 1989; Meisels & Provence, 1989). But more important than lists of tests are the criteria that should be applied to any test in order to select an appropriate instrument.

Listed in Table 1 are 4 criteria for the selection of developmental screening tests (see Meisels [1989] for a complete explanation of these criteria). Criteria for the selection of school readiness tests can also be proposed (Meisels, 1986). Table 2 presents these criteria.

Table 1

Criteria For The Selection of Developmental Screening Instruments

1. A brief procedure designed to identify children who may have a learning problem or handicapping condition that could affect their overall potential for success in school.
2. Primarily samples the domain of developmental tasks rather than the domain of specific accomplishments that indicate academic readiness.
3. Focuses on performance in a wide range of areas of development, including speech, language, cognition, perception, affect, and gross and fine motor skills.
4. Classificational data are available concerning the reliability and validity of the instrument.

(From Meisels [1989b])

Table 2

Criteria For The Selection of School Readiness Tests

1. Designed to test briefly the relative preparedness of children to participate in a specific pre-kindergarten or kindergarten program.
2. Content should be consistent with the educational values and curriculum goals of the educational program the child is about to enter.
3. Should be criterion-referenced, wherein a child's performance is indicative of a specific level or degree of mastery, rather than norm-referenced, in which a child's performance is compared to the average performance of a standardization sample.

(From Meisels [1986])

Testing is not an end in itself. It should only be used to obtain the best and most appropriate services for the greatest number of children. If the results of testing are not used—or are not used correctly—then testing should not take place. It is essential to understand how test data can be appropriately used in an educational situation to improve educational practice (see Meisels & Wasik, 1990).

G. After Testing—What?

Developmental screening tests have two principal uses: they identify children who should go on for further evaluation in order to determine if they are in need of special educational services. They can also be used to sort out children who are at-risk for school success, but do not require special education evaluation and intervention. Such children fall between the usual "OK" and "Refer" categories, and most developmental screening tests suggest that these children be rescreened after 6-8 weeks. If they remain within this area of risk, they should receive a modified classroom program designed to meet their individual needs.

School readiness tests, of course, provide different information. They are intended to facilitate curriculum planning, not to identify children needing special services. Teachers who

select readiness tests that reflect their value system and approach to curriculum planning should be able to use readiness information to make effective initial curriculum decisions and to design individuality programs.

But beyond these observations a number of questions remain in need of systematic study. For example, what alternatives to standardized testing can be developed to document student learning and to respond to needs for accountability? How can standardized tests be modified so as to be of greater use in instructional planning? What types of inservice training, ongoing supervision, and parent programs must be devised to support such innovation?

Clearly, many research questions remain to be answered. Among the most pressing are the following:

1. How wide-spread is standardized testing in kindergarten through grade 3?
2. Which tests are being used, and what are their psychometric properties?
3. Who selects the tests used in K-3? What is the basis for this selection?
4. What is the failure/retention rate in kindergarten and first grade, and how has it changed over the past 5 years?
5. At what rate are parents are holding out their children from kindergarten?
6. How many children are enrolled in extra-year programs before first grade?
7. What is the cost and funding sources of these programs?
8. What are the demographic characteristics of those children who are retained and/or enrolled in extra-year programs in terms of race, sex, socio-economic status, ethnic group, and family configuration?
9. How have curricula changed in relationship to the increased emphasis on testing?
10. What impact have tests had on teachers' classroom practices, sense of professional efficacy, and beliefs and expectations about student learning?

Answers to these questions are essential for the development of sound policy alternatives. Indeed, two general areas of inquiry are in need of further effort. First, as elaborated in this paper, educators today need sound, accurate information on the extent and use of different kinds of tests, the differential effects of standardized tests by demographic characteristics, and the impact of such tests on curricula and teachers. In addition, alternative assessment procedures must be developed that can provide a richer, more valid picture of children's educational performance and that can satisfy responsibly the nation's apparently insatiable need for accountability.

The current reliance on whole-group administered, norm-referenced tests to demonstrate accountability in the early years has, as demonstrated in this paper, contributed to more problems than it has solved. An alternative model, following the administration of a valid developmental screening instrument, would involve the use of three types of measures or procedures: 1) a criterion-referenced assessment of classroom learning, 2) a portfolio approach to documenting student progress, and 3) a systematic, standardized teacher-report form that can be used summatively to record student achievement. These three alternative measures are designed to work together, checklists indicating students' weaknesses and strengths while informing portfolio goals, as portfolio objectives inform the teachers' year-end summative report. This system poses an alternative to product-oriented standardized tests by serving as more than a mere summary of achievement. Rather than a general snapshot of academic skills at a single point in time, the ongoing evaluation process entailed by this set of alternative assessment procedures should have a positive effect on both instructional behavior and student outcomes and is intended to reflect more closely the actual goals and objectives of the classroom teacher.

A multidimensional assessment of children's progress, such as that proposed above, would have the potential for eliminating many of the problems and abuses that have accompanied early childhood testing in recent years. But such an approach, with its emphasis on encouraging teachers to make important high-stakes decisions, must be implemented with great care, supervision, and systematic research. Because of its focus on how children learn and on helping

teachers to better understand and chart individual children's styles of learning, this approach has the potential for transforming assessment information into important learning experiences. It is time to focus on educationally and developmentally appropriate assessment—assessment that takes place in the service of the child and teacher—rather than assessment that occurs at the expense of learning and at high personal cost to children, teachers, families, and the professional community at large.

II. REFERENCES

Alexander, K.L., & Entwisle, D.R. (1988). Achievement in the first 2 years of school: Patterns and processes. Monographs of the Society for Research in Child Development, 53 (2, Serial No. 218).

Alexander, K.L., Entwisle, D.R., Cadigan, D., & Pallas, A. (1987). Getting ready for first grade: Standards of deportment in home and school. Social Forces, 66, 57-84.

Alexander, K., Entwisle, D., & Thomson, M. (1987). School performance, status relations, and the structure of sentiment: Bringing the teacher back in. American Sociological Review, 52, 665-682.

American Psychological Association, American Educational Research Association, and National Council on Measurement in Education (1985). Standards for educational and psychological tests. Washington, DC: American Psychological Association.

Angoff, W.H., & Anderson, S.B. (1967). The standardization of educational psychological tests. In D. A. Payne & R. F. McMorris (Eds.), Educational and psychological measurement, (pp. 9-14). Waltham, MA: Blaisdell.

Barnes, K.E. (1982). Preschool screening: The measurement and prediction of children at-risk. Springfield, IL: Charles C Thomas.

Bear, G.G., & Modlin, P.D. (1987). Gesell's developmental testing: What purpose does it serve? Psychology in the Schools, 24, 40-44.

Boehm, A.E. (1985). Review of Brigance K & 1 Screen. In J. Mitchell, Jr. (Ed.), The ninth mental measurements yearbook (vol. 1, pp. 223-225). Lincoln, NE: University of Nebraska Press.

Boehm, A.E., & Slater, B.R. (1977). Cognitive Skills Assessment Battery. New York: Teachers College Press.

Bowen, E. (1988, April 25). Can kids flunk kindergarten? Yes, sir—especially where the law mandates tests for first grade. Time, p. 86.

Bradley, R.H. (1985). Review of Gesell School Readiness Tests. In J. Mitchell, Jr. (Ed.), The ninth mental measurements yearbook (vol. 1, pp. 609-610). Lincoln, NE: The University of Nebraska Press.

Bredekamp, S., & Shepard, L. (1989). How best to protect children from inappropriate school expectations, practices, and policies. Young Children, 44, 14-24.

Brigance, A.H. (1982). Brigance K & 1 Screen for Kindergarten and First Grade. North Billerica, MA: Curriculum Associates, Inc.

Brophy, J.E. (1983). Research on the self-fulfilling prophecy and teacher expectations. Journal of Educational Psychology, 75, 631-661.

Bruner, J.S. (1966). Towards a theory of instruction. Cambridge, MA: Harvard University Press.

Bryk, A.S., Meisels, S.J., & Markowitz, M.T. (1979). Assessing the effectiveness of open classrooms on children with special needs. In S. J. Meisels (Ed.), Special education and development: Perspectives on young children with special needs (pp. 257-296). Baltimore: University Park Press.

Cadigan, D., Entwisle, D.R., Alexander, K.L., & Pallas, A.M. (1988). First-grade retention among low achieving students: A search for significant predictors. Merrill-Palmer Quarterly, 34, 71-88.

Cahil, G. (1988, August 28). Ready or not? Kindergartens are looking for a few good kids. Marin (CA) Independent Journal, pp. E-1, E-10.

Carini, P.F. (1975). Observation and description: An alternative methodology for the investigation of human phenomena. Grand Forks, ND: University of North Dakota.

Carmody, D. (1989, May 10). Debate intensifying on screening tests before kindergarten. The New York Times, p.1, p. 14.

Charlesworth, R. (1989). "Behind" before they start? Young Children, 44, 5-13.

Chunn, E.W. (1988). Sorting black students for success and failure: The inequity of ability grouping and tracking. Urban League Review, 11, 93-106.

Cooper, D.H., & Farran, D.C. (1988). Behavioral risk factors in kindergarten. Early Childhood Research Quarterly, 3, 1-19.

Coons, P. (1987, November 29). Kindergarten: Who is ready? The Boston Sunday Globe, pp. B-77, B-79.

CTB/McGraw-Hill (1988). California Achievement Test, Grade K (Georgia Edition). Monterey, CA: author.

Cunningham, A.E. (1988). Eeny, meeny, miny, moe: Testing policy and practice in early childhood. Berkeley, CA: National Commission on Testing and Public Policy.

Darling-Hammond, L., & Wise, A.E. (1985). Beyond standardization: State standards and school improvement. The Elementary School Journal, 85, 315-336.

Dent, H.E., Mendocal, A.M., Pierce, W.D., & West, G.I. (1987). Court bans use of I.Q. tests for blacks for any purpose in California state schools: Press release by law offices of Public Advocates, Inc. San Francisco, California. The Negro Educational Review, 38, 190-191.

Detroit Dropout Prevention Collaborative (1987). Vested Interest Program: A program for children at-risk. Detroit, MI: Detroit Pubic Schools.

Durkin, D. (1987). Testing in the kindergarten. The Reading Teacher, 40, 766-770.

Engel, P. (1989, June). Assessment of kindergartener's readiness for first grade: Policies and practices of industrialized nations. Paper presented at the 1989 Annual assessment conference of the Education Commission of the States, Boulder, CO.

Entwisle, D., & Alexander, K. (1988). Factors affecting achievement test scores and marks of black and white first graders. The Elementary School Journal, 88, 450-471.

Entwisle, D.R., & Hayduk, L.A. (1981). Academic expectations and the school attainment of young children. Sociology of Education, 54, 34-50.

Fiske, E.B. (1988, April 10). America's test mania. The New York Times Spring Education Supplement, pp. 16-20.

Fiske, J. (1988, May 8). Kindergarten: The rules have changed. The Press-Enterprise (Riverside, CA), pp. B-1, B-3.

Fuchs, D., & Fuchs, L.S. (1989). Effects of examiner familiarity on black, caucasian, and Hispanic children: A meta-analysis. Exceptional Children, 55, 303-308.

Gartner, A., & Lipsky, D.K. (1987). Beyond special education: Towards a quality system for all students. Harvard Educational Review, 57, 367-395.

Gnezda, M.T., & Bolig, R. (1988). A national survey of public school testing of prekindergarten and kindergarten children. Washington, DC: National Academy of Sciences.

Gredler, G.R. (1978). A look at some important factors in assessing readiness for school. Journal of Learning Disabilities, 11, 284-290.

Haines, J., Ames, L.B., & Gillespie, C. (1980). The Gesell Preschool Test Manual. Lumberville, PA: Modern Learning Press.

Hamilton, J.L., & Swan, W.W. (1981). Measurement references in the assessment of preschool handicapped children. Topics in Early Childhood Special Education, 1, 41-48.

Hechinger, F.M. (1988, September 14). Repeating kindergarten: Does it hurt more than it helps? New York Times, p. 24.

Hein, G.E. (1979). Evaluation in open education: Emergence of a qualitative methodology. In S. J. Meisels (Ed.), Special education and development: Perspectives on young children with special needs (pp. 231-250). Baltimore, MD: University Park Press.

Heller, K., Holtzman, W., & Messick, S. (Eds.) (1982). Placing children in special education: A strategy for equality. Washington, DC: National Academy Press.

Holmes, C.T., & Matthews, K.M. (1984). The effects of nonpromotion on elementary and junior high school pupils: A meta-analysis. Review of educational research, 54, 225-236.

Ilg, F.L., & Ames, L.B. (1972). School Readiness. New York: Harper & Row, 1982.

Kaufman, N.L. (1985). Review of Gesell Preschool Test. In J. Mitchell, Jr. (Ed.), The ninth mental measurements yearbook (vol. 1, pp. 607-608). Lincoln, NE: The University of Nebraska Press.

Kelleghan, T., Madaus, G.F., & Airasian, P.W. (1982). The effects of standardized testing. Boston, MA: Kluwer-Nijhoff Publishing.

Koretz, D. (1988). Arriving in Lake Wobegon: Are standardized tests exaggerating achievement and distorting instruction? American Education, Summer, 8-15; 46-52.

Lichtenstein, R. (1980). The Minneapolis Preschool Screening Inventory. Minneapolis: Minneapolis Public Schools.

Lichtenstein, R., & Ireton, H. (1984). Preschool screening: Identifying young children with developmental and educational problems. Orlando, FL: Grune & Stratton.

Madaus, G.F. (1988). The influence of testing on the curriculum. In L. N. Tanner (Ed.), Critical issues in curriculum, 87th Yearbook of the National Society for the Study of Education (pp. 83-121). Chicago: University of Chicago Press.

May, D.C., & Campbell, R.M. (1981). Readiness for learning: Assumptions and realities. Theory Into Practice, 20, 130-134.

May, D.C., & Welch, E.L. (1984a). The effects of developmental placement and early retention on children's later scores on standardized tests. Psychology in the Schools, 21, 381-385.

May, D.C., & Welch, E.L. (1984b). Developmental placement: Does it prevent future learning problems? Journal of Learning Disabilities, 17, 338-341.

McCarthy, D. (1972). McCarthy Scales of Children's Abilities. New York: Psychological Corporation.

Meisels, S.J. (1984). Prediction, prevention, and developmental screening in the EPSDT program. In H. W. Stevenson & A. G. Siegel (Eds.), Child development research and social policy, (pp. 267-317). Chicago: University of Chicago Press.

Meisels, S.J. (1986). Testing four- and five-year olds. Educational Leadership, 44, 90-92.

Meisels, S.J. (1987a). Uses and abuses of developmental screening and school readiness testing. Young Children, 42, 4-6; 68-73.

Meisels, S.J. (1987b). Using criterion-referenced assessment data to measure the progress of handicapped children in early intervention programs. In G. Gasto, S. Ascione, & M. Salehi (Eds.), Perspectives in infancy and early childhood, (pp. 59-64). Logan, UT: DCHP Press.

Meisels, S.J. (1988). Developmental screening in early childhood: The interaction of research and social policy. In L. Breslow, J. E. Fielding, & L. B. Lave (Eds.), Annual Review of Public Health (vol. 9, pp. 527-550). Palo Alto, CA: Annual Reviews.

Meisels, S.J. (1989a). Developmental screening in early childhood: A guide. Third edition. Washington, DC: National Association for the Education of Young Children.

Meisels, S.J. (1989b). Can developmental screening tests identify children who are developmentally at risk? Pediatrics, 83, 578-583.

Meisels, S.J. (1989c). High stakes testing in kindergarten. Educational Leadership, 46, 16-22.

Meisels, S.J., Harbin, G., Modigliani, K., & Olson, K. (1988). Formulating optimal state early childhood intervention policies. Exceptional Children, 55, 159-165.

Meisels, S.J., & Provence, S. (1989). Screening and assessment: Guidelines for identifying young disabled and developmentally vulnerable children and their families. Washington, DC: National Center for Clinical Infant Programs.

Meisels, S.J., & Wasik, B.A. (1990). Who should be served? Identifying children in need of early intervention. In S. J. Meisels & J. P. Shonkoff (Eds.), Handbook of early childhood intervention (pp. 605-632). New York: Cambridge University Press.

Meisels, S.J., & Wiske, M.S. (1983). The Early Screening Inventory (second edition). New York: Teachers College Press.

Naglieri, J.A. (1985). Review of Gesell Preschool Tests. In J. Mitchell, Jr. (Ed.), The ninth mental measurement yearbook (vol. 1, pp. 608-609). Lincoln, NE: The University of Nebraska Press.

National Association for the Education of Young Children (1988). Position statement on standardized testing of young children 3 through 8 years of age. Young Children, 43, 42-47.

National Association of Early Childhood Specialists in State Departments of Education (1987). Unacceptable trends in kindergarten entry and placement: A position statement. Lincoln, NE: author.

National Association of State Boards of Education (1988). Right from the start: The report of the NASBE Task Force on early childhood education. Alexandria, VA: author.

National Commission on Excellence in Education (1983). A nation at risk: The imperative for educational reform. Washington, DC: U.S. Government Printing Office.

Nurss, J.R., & McGauvran, M.E. (1976). Metropolitan Readiness Tests. New York: Harcourt, Brace, Jovanovich.

Ordovensky, P. (1989, June 14). Repeating a grade may drive kids to drop out. USA Today, p. D-1.

Pallas, A., Entwisle, D., Alexander, K., & Cadigan, D. (1987). Children who do exceptionally well in first grade. Sociology of Education, 60, 257-271.

Parsons, J.E., Adler, T.F., & Kaczala, C.M. (1982). Socialization of achievement attitudes and beliefs: Parental influences. Child Development, 53, 310-321.

Parsons, J.E., Kaczala, C.M., & Meece, J.L. (1982). Socialization of achievement attitudes and beliefs: Classroom influences. Child Development, 53, 322-339.

Plummer, D.L., Lineberger, M.H., & Graziano, W.G. (1986). The academic and social consequences of grade retention: A convergent analysis. In L. G. Katz (Ed.), Current topics in early childhood education, (vol. 6, pp. 224-252). Norwood, NJ: Ablex Publishing Co.

Popham, W.J., & Husek, T.R. (1969). Implications of criterion referenced measures. Journal of Educational Measurement, 6, 1-9.

Putka, G. (1988, July 6). Tense tots: Some schools press so hard kids become stressed and fearful. The Wall Street Journal, pp. 1, 6-7.

Riley, S., Carter, P., Cummings, C., Firestone, J., Flynn, C., Javid, S., Ruiter, D. (1988). Survey results: Early childhood programming. Paper presented at state kindergarten conference, Flint, MI, September, 1988.

Rist, R.D. (1970). Student social class and teacher expectations: The self-fulfilling prophecy in ghetto education. Harvard Educational Review, 40, 411-451.

Robinson, J.H., & Kovacevich, D.A. (1984). The Brigance Inventories. In D. J. Keyser & R. C. Sweetland (Eds.), Test critiques (vol. 1, pp. 79-98). Kansas City, MO: Test Corporation of America.

Salmon-Cox, L. (1981). Teachers and standardized achievement tests: What's really happening? Phi Delta Kappan, 62, 631-633.

Salzer, R. (1986). Why not assume they're all gifted rather than handicapped? Educational Leadership, 44, 74-77.

Shepard, L.A. (1989). Why we need better assessments. Educational Leadership, 46, 35-40.

Shepard, L.A., & Smith, M.L. (1985). Boulder Valley Kindergarten Study: Retention practices and retention effects. Boulder, CO: Boulder Valley Public Schools.

Shepard, L.A., & Smith, M.L. (1986). Synthesis of research on school readiness and kindergarten retention. Educational Leadership, 44, 78-86.

Shepard, L.A., & Smith, M.L. (1987). Effects of kindergarten retention at the end of first grade. Psychology in the Schools, 24, 346-357.

Shepard, L.A., & Smith, M.L. (1988). Escalating academic demand in kindergarten: Counterproductive policies. Elementary School Journal, 89, 135-145.

Smith, M.L., & Shepard, L.A. (1987). What doesn't work: Explaining policies of retention in the early grades. Phi Delta Kappan, 69, 129-134.

Taylor, O.L., & Lee, D.L. (1987). Standardized tests and African-American children: Communication and language issues. The Negro Educational Review, 38, 67-80.

Teale, W.H., & Sulzby, E. (1986). Emergent literacy as a perspective for examining how young children become writers and readers. In W. H. Teale, & E. Sulzby (Eds.), Emergent literacy: Writing and reading (pp. vii-xxv). Norwood, NJ: Ablex.

Wodtke, K.H., Harper, F., Schommer, M., and Brunelli, P. (1985). Social context effects in early school testing: An observational study of the testing process. Paper presented at American Educational Research Association, Chicago, Illinois, April, 1985.

Wood, C., Powell, S., & Knight, R.C. (1984). Predicting school readiness: The validity of developmental age. Journal of Learning Disabilities, 17, 8-11.

Yagelski, R.P. (1988, January 17). Fears for a son going into a test-crazy world. The Boston Globe, pp. A-44, A-48.